Social Media-Driven Credit Scoring: the Predictive Value of Social Structures

Research-in-Progress

Tianhui Tan
Department of Information Systems,
National University of Singapore
15 Computing Drive, Singapore 117418
tianhui.tan@u.nus.edu

Tuan Q. Phan
Department of Information Systems,
National University of Singapore
15 Computing Drive, Singapore 117418
disptq@nus.edu.sg

Abstract

While emerging economies have seen an explosion of social network site (SNS) adoption, these countries lack sophisticated credit scoring system or credit bureaus to predict creditworthiness of individuals. In this paper, we propose an SNS-based credit scoring method for micro loans using largescale observational data. We show empirical evidence that by incorporating social network metrics, we can improve the repayment prediction by 18%. Next, to better harness the combination of borrowers and non-borrowers, we implement graph-based prediction using the semi-supervised learning method. At current stage, by only leveraging on social ties, the prediction performance looks promising with an accuracy of 60%. We believe that although lending to the poor without incurring high default rates is challenging, the SNS-based methods can be effective for developing countries that face the "cold-start" problem.

Keywords: Social networks, Credit scoring, Microfinance, Logistic regression, Semi-supervised learning

Introduction

Financial inclusion has been widely recognized to be critical in reducing poverty and boosting prosperity. However, in 2014, 38 percent of adults in the world still remain unbanked (Demirgüç-Kunt, et al., 2015). Especially in underdeveloped countries, lack of a sophisticated credit reporting system or credit bureaus makes creditworthiness assessment challenging, giving rise to financial exclusion. In the hopes of boosting credit growth, other forms of credit and credit assessment have emerged such as microfinance, which do not rely on financial histories and physical collaterals. In these cases, social capital becomes a popular substitute. For example, Grameen Bank makes use of social capital to ensure the repayment of group borrowers (Yunus, 2007). Recently, Kiva launched a "social underwriting" program in U.S for financially excluded borrowers to establish the creditworthiness using networks and character (Simon, 2015). On the other hand, the explosive growth and penetration of social network site (SNS) provides plentiful social information. Through leveraging borrowers' information on the online platforms, some startups create new ways of credit scoring. For example, Lenddo, takes this initiative of SNS-based scoring to derive credit by appraising applicants' social media information (Groenfeldt, 2015). In China, Wecash develops credit profiles by synthesizing borrowers' information from various sources including emails and messegers such as Tencent QQ (Weinland and Robertson, 2014). Yet, with this plethora of social information, there is little guidance and method to effectively leverage it in the context of credit scoring.

Here, we explore a new method based on social structures for credit scoring using SNS data. We demonstrate its effectiveness on a dataset of 3661 microfinance loans from a popular SNS-based service.

SNS-based credit scoring relies on the notion of homophily - “birds of a feather tend to flock together” (Lazarsfeld and Merton, 1954; McPherson, et al., 2001). In other words, by leveraging social ties, the characteristics such as creditworthiness could be inferred. However, exploiting SNS data to predict creditworthiness is a non-trivial task. In social networks, borrowers freely connect to other borrowers with known creditworthiness and non-borrowers without. Borrower’s attributes-based credit scoring methods such as regression and classification focus on users’ preferences, likes, and characteristics in isolation. However, social structures tend to be omitted in these models. Yet, they can be captured using relevant peer-to-peer interactions and social network metrics. That is, incorporating social structure can capture additional information and latent variables.

Furthermore, the role of non-borrowers tends to be overlooked in traditional methods. Although non-borrowers themselves do not have repayment records, they carry and transmit valuable information by connecting to borrowers in the networks. Hence, a more plausible approach is to analyze the network graph as a whole. Learning through the combination of borrowers and non-borrowers would greatly enhance our understanding (Zhu & Goldberg, 2009). As a result, our approach is to enhance credit scoring using an SNS-based approach which incorporates both borrowers and non-borrowers networks and attributes.

To test our method, we obtain anonymized backend data from a company which offers microfinance loans in a Southeast Asian country. Without banking history, financially underprivileged individuals are likely to be underserved. However, this lending company encourages borrowers to share their SNS accounts to obtain credit. Our data consists of both loan repayment record from the company, and corresponding borrower’s Facebook profile and her full history of interactions on Facebook. As a baseline, we start with attributes-based models using users’ preferences, namely stated interests (eg. Fanpage “likes”) and group participations (eg. Facebook groups) and implement logit regression and random forest to explore the predictability of SNS data. We show that SNS-based predictors alone improve the prediction by more than 18%, as compared to the model with only demographics predictors. We then introduce our graph-based prediction. At the current stage, we only incorporate an edge attribute - tie strength as measured by Facebook interactions including posts and messages from the borrowers to both other borrowers and non-borrowers. Our tests show that the weighted social ties alone correctly predict the repayment for more than 60% borrowers. Overall, we show that credit scoring can be improved by leveraging social networks and structure.

Our study makes a few important contributions. To the best of our knowledge, we are among the first to empirically predict repayment based on social networks. By building social network-based creditworthiness assessment, our study adds value to the finance literature and extends existing social network studies through utilizing a large scale, objective network data to investigate the effects of social networks (e.g. Eagle, et al., 2009; Yoganarasimhan, 2012; Fang, et al., 2013). More importantly, from a pragmatic view, we show that by considering network ties, we can improve our predictions over traditional, non-network based methods. For traditional finance practitioners such as banks, analyzing social network data will help develop better credit scoring as it carries important unobservable information about borrowers. In underdeveloped areas where people have limited credit access due to lack of finance background or collateral, practitioners could consider about leveraging SNS data to calculate the credit scores. The SNS-based credit scores will help alleviate financial exclusion.

Background

Methods in Credit Scoring

Consumer credit has existed since Babylonian times (Lewis, 1992), but only after Durand’s (1941) research on discriminating good and bad loans, credit scoring was realized as a technique to assist organizations in making credit decisions (Thomas, et al., 2005). Although Durand (1941) first adopted statistical methods to select credit applicants, the fact that credit scoring was being largely applied in credit industry practices in a predictive and pragmatic manner, should be largely attributed to the consultancy, Fair, Isaac & Company Inc., in the early 50s (Poon, 2007). With the rapid development of computer technology, the credit scoring systems were widely spreading in early 1960s (Capon, 1982). The credit scoring is a set of decision models assessing the risk in lending to a particular borrower at the moment of underwriting (Mays, 2001; Thomas, et al., 2002).

The original models dated from the 1950s were mainly statistical discrimination and classification (Thomas, et al., 2002). Wiginton (1980) was the first to adopt logistic regression and its prediction was reasonably accurate. Others compared among multiple methods (e.g. Srinivasan and Kim, 1987; Desai, et al., 1997; Hand and Henley, 1997; West, 2000; Baesens, et al., 2003) and logistic regression was found to be a simple and efficient classification method for credit scoring. In addition, advanced credit screening and granting methods were developed, including neural network (Jensen, 1992), nonparametric approach such as the nearest neighbor method (Chatterjee and Barcun, 1970), support vector machine (Huang, et al., 2007), genetic programming (Ong, et al., 2005) and other hybrid approaches (Hsieh, 2005; Lee and Chen, 2005). However, the results for comparing prediction accuracy of different methods were mixed (e.g. Desai, et al., 1997; West, 2000; Yobas, et al., 2000; Baesens, et al., 2003). In other words, there was not a general model superior to other models. All the models could perform well, depending on the data structure and research context (Hand, 1997).

The credit score has been built mainly based on credit bureau data using past repayment history as well as loan application data such as collateral and employment information (Rosenberg and Gleit, 1994; Thomas, et al., 2005). Credit bureaus would collect borrower's performances from various lenders and generate credit reports. For example, the popular FICO has been a credit bureau score. The mortgage scoring system incorporated predictors from all resources including borrowers' income and collateral to predict for a specific mortgage loan (Mays, 2001).

As Orgler (1970) commented in early years, although predictors varied somewhat among different models, they were all based on borrower's details such as credit record and finance background. There were few exceptions such as the student loan, which required no prior experience and collateral (Flint, 1997). Those government-initiated student loan programs conducted risk evaluation by surveying details on family background, expected job and income (Dynarski, 1994). For other cases, lenders would generally rely on credit bureau data and/or loan application data to assess borrowers' creditworthiness. In other words, without these evidences, it was difficult for borrowers to get credit. For example, people in underdeveloped countries without credit bureaus such as Nepal (World Bank, 2015), had limited access to credit, leading to the issue of financial exclusion. Hence it became imperative to innovate credit scoring methods.

Social Networks and Microfinance

Social Capital in Microfinance Group Lending

Microfinance was first introduced in 1980s to ameliorate financial exclusion (Brau and Woller, 2004) by allowing for joint liability lending to a group of borrowers. It was considered a breakthrough strategy in economic development (Ahlin and Townsend, 2007) and became a popular method to reach underdeveloped countries. Lending to the poor without incurring high default rates was a difficult task. Group lending harnessed social capital and took advantage of what it coined as *social collateral* to ensure repayments. The underlying mechanisms included risk sharing, monitoring, trust, altruism or social pressure (e.g. Besley and Coate, 1995; Karlan, 2005; Karlan, 2007; Feigenberg, et al., 2010).

However, most of these studies investigated social capital through self-reported surveys. For instance, Karlan (2007) measured social connections in terms of geographic proximity and cultural similarity through survey questions. As a result, the extent of social interactions in the networked environment was limited. Hence, these studies only captured networks within a small and well-bounded group, rather than the borrower's social network in everyday life (Wasserman and Faust, 1994). Moreover, it remained uncovered whether social capital would exert influence on individual loans. Hence, we provide a different approach to leverage social capital in the context of individual loans and explore the social network effects using real-world SNS data.

Social Network-based Credit Scoring

Previous researchers have noted that individuals who were similar tend to form relationships – giving rise to the idiom “birds of a feather tend to flock together,” or homophily in the academic literature (Lazarsfeld and Merton, 1954; McPherson, et al., 2001; Currarini, et al., 2009). The impact of homophily has been well studied in various areas such as product adoption (Centola, 2011) and employment (Fernandez, et al., 2000). Through the leverage of homophily, borrowers’ creditworthiness can be assessed by analyzing their social structures. In other words, a trustworthy person’s friend is more likely to be trustworthy.

However, the existing studies often captured homophily using observable factors (Aral, et al., 2009) while latent homophily due to unobservable personal traits remained a challenge (Shalizi and Thomas, 2011). Although we are able to capture observable homophily based on the information disclosed by borrowers on their public SNS profiles, we argue that we can increase the predictability by capturing unobservable homophily through online social connections. That is, one shares similar traits with the one she interacts, but these traits may not be openly stated in their online profiles. This latent homophily may take the form of offline activities which are not directly recorded by the online platforms nor observed by finance practitioners. Instead, these latent attributes can be and are captured through social ties and peer-to-peer interactions on the site. Thus, social network data can be a promising source to predict loan repayment. To the best of our knowledge, there has been scant literature exploring whether social network information could help with credit scoring for individual loans. Wei et al.(2016) offers some insights by developing an analytical model to leverage social network for credit scoring. However, there is still a lack of empirical evidence.

When finance services find it difficult to assess creditworthiness for borrowers with no finance background, online social networks might provide a efficient data source and offer useful information about the borrowers. Hence, in this study, we aim to develop a method for credit scoring based on borrowers’ social networks and demonstrate that *considering network-based data will significantly improve the predictive power of ego-based predictions*.

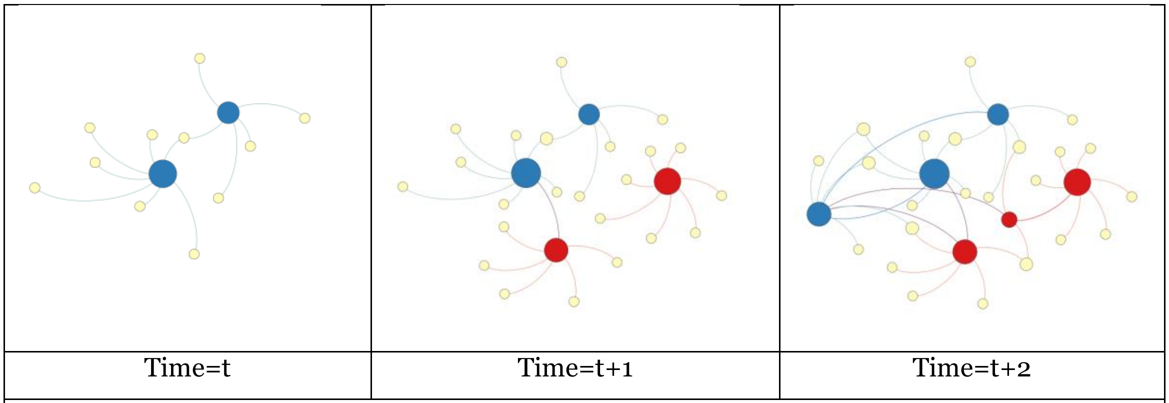
Data Description

We obtain anonymized backend data from a company which offers microfinance loans mainly towards the middle class in a Southeast Asian country. The country has a annual per capita GDP of less than USD 3000, categorized as a lower middle income economy. It does not have existing credit scoring bureaus, thus without collateral, people have been faced with difficulties to get credit. These people have low access to and usage of financial services, and only 12% of adults in the country have a loan from a formal financial institution in 2014. On the other hand, the country has fast growing Internet population, experiencing more than 500% growth in the last five years. People quickly adopt to technical innovations such as social media and messenger apps. Specifically, in 2015, the penetration of Facebook has reached more than 42%. Hence, the lending company can leverage social media to calculate credit scores. Instead of collaterals and financial histories, borrowers are encouraged to share their SNS (e.g. Facebook) accounts with the lender to obtain better credit scores. On average, the lender offers small loans equivalent of USD 300 for a duration of 3-6 months.

To start the loan application process, the borrower will register herself on the lender’s website. Then the borrower connects the profile with her social networks and grants the lender to access her SNS information. A “score” will be shown to the borrower, which depends on the information provided in the profile. The score is then used to calculate the interest, or “cost,” of the loan to the borrower. The process is completely online with minimal paperwork. Being granted access to the borrower’s social media data, the lender is able to analyze the borrower’s social activities with direct connections from her SNS accounts. Furthermore, the borrower’s friends are not aware of her defaulting - she can act without fear of retribution. Social pressure is no longer leveraged in our context. As a result, creditworthiness prediction is crucial to the viability of the company.

Our anonymized data consists of two components: first, a sample of micro-loans from 2011 to 2014 with repayment records; second, the corresponding borrowers’ Facebook profiles and full histories of time-stamped Facebook activities including posts, messages, photos, etc, till 2014. We identify the individuals who have ever borrowed from the lender as borrowers (i.e. ego) and their direct friends (i.e. alters). If the borrower repays the loan, she is categorized as a good borrower; otherwise, she is a bad borrower. At the current stage, we focus on first-time borrowers.

Although the data is large, we are faced with a few challenges. The first caveat is that, due to privacy concerns, for each borrower (i.e. ego), the lender only accesses her social interactions with direct connections (i.e. alters). Thus, we can neither capture the complete social network, nor extend to networks of alters. That is, we can only observe the ego-centric network. However, the ego can connect to other borrowers, the "prior borrowers," who have applied loans earlier than the ego. For prior borrowers, the lender is able to obtain their interactions with direct friends (see Figure 1). Moreover, the repayment outcomes of these prior borrowers are also known by the lender. Therefore, ego's network structure can be further reduced to creditworthy-only substructure (involving only ego and prior borrowers who repaid their loans) and creditworthless-only substructure (involving only ego and prior borrowers who defaulted). The network connections in these two substructures are fully captured. More importantly, lenders making credit decisions (unlike platform owners) may not know the complete network structure either. Thus, an exploratory study of easy-to-collect local and incomplete network is certainly managerially relevant. Second, we solely rely on the social network information (from the SNS) to predict repayments. However, we believe that if SNS data alone can produce accurate enough prediction, SNS backed tactics will at least provide effective data source for real-world practices, which is consistent with the pragmatism idea of credit scoring (Thomas, 2005).



Time=t

Time=t+1

Time=t+2

Figure 1. Illustration on SNS Data Collection from Lender's View

Note: blue nodes indicate good borrowers; red nodes indicate bad borrowers; yellow nodes indicate alters; edges indicate there exist interactions between two nodes

Methods and Preliminary Results

Attributes-based Prediction

The dependent variable $egoRepaid_i$ is a binary, equaling 1 for repayment and 0 for default. The predictors we derive can be divided into the following three sets, i.e. demographics, preferences and networks. We obtain the demographics from borrowers' Facebook profile pages, including age, gender, marital status, religion, education and location.

We obtain the borrower's preferences through her interests (e.g. "Taylor Swift") and groups (e.g. "I Love XXX City") indicated on Facebook profile pages. This set of variables is used to capture observable homophily. We identify 95201 distinct Facebook groups and interests. For each of the group and interest, we construct a dummy variable. These personal preference indicators may contain hints on individual traits or personal background that relate to creditworthiness.

We obtain the borrower's network information through her social interactions on Facebook. We extract all of the ego's timestamped Facebook activities prior to her loan application and aggregate across communication channels such as messages and posts. Relevant social network metrics are calculated (Wasserman and Faust, 1994; Barrat, et al., 2004). The first metric is the $indegree_i$ which is the number of direct friends or alters that initiate interactions to ego. Here indegree is preferred (over outdegree), because the ego has less control over friendship requests. It is harder for the ego to manipulate indegree.

We also calculate $degree_i$, which is the total number of alters ever interacted with ego. Next, we extend the definition of degree to include the actual strength of the tie (which is estimated as the number of interactions between two nodes), defined as $strength_i$. It is a sum of the strength for all the ego's social ties, so it quantifies ego's total number of interactions with alters. Thus, $strength_i$ divided by $degree_i$ becomes the average tie strength for the ego, denoted by avg$Strength_i$. Last, we calculate the triadic closure of ego's local network, defined as $weightedTransitivity_i = \frac{1}{strength_i(degree_i-1)}\sum_{j,h}\frac{(w_{ij}+w_{ih})}{2}e_{ij}e_{ih}e_{jh}$ where e_{ij}, e_{ih}, e_{jh} are connections for nodes (i, j), nodes (i, h) and nodes (j, h), and w_{ij}, w_{ih} indicate the numbers of interactions for nodes (i, j) and nodes (i, h) respectively (Barrat, et al., 2004). To summarize, as stated in the above data description section, we derive the creditworthy-only (creditworthless-only) subnetworks from the ego's overall network. For creditworthy-only (creditworthless-only) sub-structure, we calculate three metrics $goodIndegree_i$ $(badIndegree_i)$, $goodAvgStrength_i$ $(badAvgStrength_i)$ and $goodWeightedTransitivity_i$ $(badWeightedTransitivity_i)$, to capture the two fundamental concepts – connectivity and clustering in network. For the overall structure, we only calculate $totalDegree_i$ and $totalAvgStrength_i$, because transitivity will be bias if some alters' connectivity is unknown.

To summarize, we obtain a sample of 1047 borrowers with demographics, preference and network information available. In this sample, the repayment rate is 87.5%. Table 1 shows the summary statistics which also implies the sparsity of borrowers in the network.

Table 1. Summary Statistics

Statistic	Mean	St. Dev.	Min	Max
egoGood	0.875	0.331	0	1
female	0.554	0.497	0	1
age	29.549	5.975	18	59
hasRel	0.350	0.477	0	1
loc_Luzon	0.243	0.429	0	1
loc_Mindanao	0.035	0.185	0	1
loc_Visayas	0.080	0.272	0	1
loc_Capital	0.607	0.489	0	1
loc_Overseas	0.034	0.182	0	1
eduLevel	2.056	0.371	1	3
status_relationship	0.264	0.441	0	1
status_single	0.380	0.486	0	1
status_married	0.356	0.479	0	1
goodIndegree	0.565	1.142	0	18
goodAvgStrength	37.313	165.115	0.000	2,989.400
goodWeightedTransitivity	0.058	0.219	0.000	1.000
badIndegree	0.724	1.298	0	10
badAvgStrength	23.541	91.707	0.000	1,450.000
badWeightedTransitivity	0.061	0.216	0.000	1.000
totalDegree	609.595	475.294	1	4,168
totalAvgStrength	20.601	22.959	1.000	213.162

Logistic regression is among the most classic and popular methods in credit scoring. Before we apply logit regression, we select a smaller set of preference dummies using random forest (Breiman, 2001). Random forest is suitable for feature selection and it avoids issues like overfitting or instability to small changes in the learning data (Strobl, et al., 2009). We start with the baseline logit model using only demographics to estimate the probability that $egoRepaid_i$ = 1 (Hosmer and Lemeshow, 2004). Then we gradually add in the set of preference predictors and the set of network predictors. In addition, we also adopt random forest for repayment prediction, following the same steps as in logit regression.

To evaluate how incorporating social network data affects predicting creditworthiness, we calculate AUC (i.e. Area under the ROC curve). It measures the probability that a randomly chosen 'creditworthy' borrower will be predicted to have a higher repayment probability than a randomly chosen 'creditworthless' borrower (Fawcett, 2006). The AUC is considered as the standard measure of the discrimination power of a classifier (Huang and Ling, 2005), and it is also known to be independent with respect to class distribution and sidestep misclassification cost (Baesens, et al., 2003). The higher AUC implies greater predictive power. We compare the predictability using a 10-fold cross validation in Table 2 below.

Table 2. Predictability Comparison

Predictors	Logit Model	Random Forest
demographics	0.5229388	0.5523359
+ set of preference dummies	0.5433915	0.5612787
+ set of network metrics	0.6185653	0.6540369

According to Table 2, SNS data alone improves the predictability by more than 18%. Thus we show some evidence that social network data could be a promising source of creditworthiness predictors. We believe that more accurate credit scoring can be achieved by leveraging social networks.

Graph-based Prediction

In the attributes-based models, effects of non-borrowers are only counted in two predictors – totalDegree and totalAvgStrength. We largely ignore their roles in the overall network structure. In fact, in our dataset, borrowers are relatively scarce, and non-borrowers act as bridges between borrowers, contributing to the connectedness of network. Borrowers can be indirectly connected via non-borrowers. Moreover, although non-borrowers do not have loan repayment information, they carry and transmit valuable credit information of their friends. Thus, overlooking the role of non-borrowers will lead to underestimation of the predictive power of social structures. In order to take full advantage of the combination of borrowers and non-borrowers in the network, we adopt a graph-based semi-supervised classification method (Zhu, et al., 2003). Using this method, we directly analyze the network structure as a graph, rather than focus on metrics derived from the graph.

At current stage, the graph-based credit scoring is solely based on social ties. Since ties represent many characteristics such as social relationship (eg. "friend", "parent", "teacher", etc..) and vary in strength (Granovetter, 1973; Van den Bulte and Wuyts, 2007; Chen, et al., 2016), we construct a weighted graph for each ego. The tie strength is computed based on number of SNS interactions prior to ego's loan application. Furthermore, we assume that tie strength indicates the similarity between two nodes (Zhu, et al., 2003). The predicted creditworthiness of ego is the weighted average among all her direct connections' perceived creditworthiness including non-borrowers.

Following the specification by Zhu, et al. (2003), we categorize all nodes in the graph into either labeled nodes (i.e. prior borrowers) or unlabeled nodes (i.e. non-borrowers or borrowers who borrow loans after ego's application). We add labels for the prior borrowers according to their repayment records. In other words, we classify labelled and unlabeled nodes based on whether the company is able to observe the individual's creditworthiness label or not at the moment of underwriting ego's loan application. The ego is categorized as an unlabeled node. L is the set of labelled nodes and U is the set of unlabeled nodes. Built upon the notion of homophily, our objective is to minimize the disagreement between any connected nodes, $\sum_{i\sim j} w_{ij}(f(i) - f(j))^2$, by solving for f which is a real-valued harmonic function. We define

$f(i) = y_i$ on the labeled node $i \in L$, and y_i denotes her label which is fixed. For all the unlabeled nodes, the function value of f will be estimated as the weighted average of direct alters' values of f (see Figure 2).

```
Step 1: Set $f(i) = y_i$ for $i \in L$ and $f(j) = 0$ for $j \in U$
Step 2: Set $f(j) = \frac{\sum_{j\sim k} w_{jk} f(k)}{\sum_{j\sim k} w_{jk}}$ for $\forall j \in U$ where $k \in L + U$, $w_{jk}$ is the tie strength between $j$ and $k$, and $f(L)$ is fixed
Step 3: Repeat Step 2 until convergence
```

Figure 2. Semi-supervised Learning Algorithm

Considering only social ties, we apply the above algorithm to a larger sample of 3661 borrowers with repayment rate of 46.52%. In total, we construct 3661 weighted graphs. On average, the network size is 101,193 and the ratio of borrowers over non-borrowers in these networks is around 0.2, suggesting the sparsity of labeled nodes and necessity of learning unlabeled nodes. By applying semi-supervised learning algorithm to predict ego's creditworthiness, we obtain an AUC of 0.6046531. Our preliminary result shows promising results.

Concluding Remarks

In this study, we adopt both classic credit scoring methods and graph-based semi-supervised learning to explore SNS-based creditworthiness assessment. We highlight the important role of non-borrowers in a network context. The preliminary results from our study are promising and help us make several important contributions. Practically, we demonstrate how social networks can be used for credit scoring. By exploiting network structures, social network-based credit scoring can help financial services in decision-making. For finance practitioners, analyzing social network data will improve credit scoring as it carries important unobservable information about the borrowers. Based on this, financial services can be offered to individuals without collateral or credit history. Theoretically, we add to the social network literature and the emerging microfinance literature by empirically uncovering associations between the social network and loan repayment behavior, and by investigating potential predictors for individual micro-loans outcomes.

However, our study in its current form has several limitations. There are other factors that may exert influence on repayment behavior, such as economic conditions and loan amounts. Due to data constraint, we do not include these factors. We also suffer from the sparsity of network data, a common problem in the research on networks. As next steps, we will improve the regression model by taking network autocorrelation into account (e.g. Lee, 2007; Bramoullé, et. al., 2009). We will also enhance the graph-based prediction by developing a method which better suits our context. Specifically, we aim at a graph-based semi-supervised learning on a network where labelled and unlabeled nodes have different levels of information completeness, in terms of network structure and node attributes. This also makes our study differentiable from another stream of studies about machine learning on social structures (e.g. Pennacchiotti and Popescu, 2011; Wu, et al., 2013). Hence, we aim to provide generalizable methodologic implication to address the realistic data issues. Moreover, the scalability and stability of prediction will be tested. Last but not least, we would like to explore individual differences in SNS-activity behaviors between creditworthy and creditworthless borrowers.

References

Ahlin, C., & Townsend, R. M. 2007. Using Repayment Data to Test Across Models of Joint Liability Lending. *The Economic Journal* (117:517), F11-F51.

Aral, S., Muchnik, L., & Sundararajan, A. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, (106:51), pp. 21544-21549.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* (54:6), pp. 627-635.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. 2004. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* (101:11), pp. 3747-3752.

Besley, T., &Coate, S. 1995. Group lending, repayment incentives and social collateral. *Journal of development economics* (46:1), pp. 1-18.

Bramoullé, Y., Djebbari, H., & Fortin, B. 2009. Identification of peer effects through social networks. *Journal of econometrics* (150:1), pp. 41-55.

Brau, J. C., & Woller, G. M. 2004. Microfinance: A comprehensive review of the existing literature. *Journal of Entrepreneurial Finance* (9:1),pp. 1-27.

Breiman, L. 2001. Random forests. *Machine learning* (45:1), pp.5-32.

Capon, N. 1982. Credit scoring systems: A critical analysis. *The Journal of Marketing* (46:2), pp. 82-91.

Centola, D. 2011. An experimental study of homophily in the adoption of health behavior. *Science* (334:6060), pp.1269-1272.

Chatterjee, S., & Barcun, S. 1970. A nonparametric approach to credit screening. *Journal of the American statistical Association* (65:329), pp.150-154.

Chen, X., van der Lans, R. & Phan, T. 2016. Uncovering the Importance of Relationship Characteristics in Social Networks: Implications for Seeding Strategies. *Journal of Marketing Research*, forthcoming.

Currarini, S., Jackson, M. O., & Pin, P. 2009. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* (77:4), pp.1003-1045.

Demirgüç-Kunt, A., Klapper, L. F., Singer, D., & Van Oudheusden, P. 2015. *The Global Findex Database 2014: measuring financial inclusion around the world*. World Bank Policy Research Working Paper, (7255).

Desai, V. S., Conway, D. G., Crook, J. N., & OVERSTREET, G. A. 1997. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics* (8:4), pp.323-346.

Durand, D. 1941. *Risk elements in consumer instalment financing*. NBER Books.

Dynarski, M. 1994. Who defaults on student loans? findings from the national postsecondary student aid study. *Economics of education review* (13:1), pp. 55-68.

Eagle, N., Pentland, A. S., &Lazer, D. 2009.Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* (106:36), pp.15274-15278.

Fang, X., Hu, P. J. H., Li, Z., & Tsai, W. 2013.Predicting adoption probabilities in social networks. *Information Systems Research* (24:1), pp.128-145.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters* (27:8), pp.861-874.

Feigenberg, B., Field, E. M., &Pande, R. 2010. *Building social capital through microfinance* (No. w16018).National Bureau of Economic Research.

Fernandez, R. M., Castilla, E. J., & Moore, P. 2000. Social capital at work: Networks and employment at a phone center. *American journal of sociology* (105:5), pp1288-1356.

Flint, T. A. 1997. Predicting student loan defaults. *Journal of Higher Education* (68:3), pp.322-354.

Granovetter, M. S. 1973. The strength of weak ties. *American journal of sociology* (78:6), pp.1360-1380.

Groenfeldt, T. January 29, 2015. *Lenddo Creates Credit Scores Using Social Media*. *Forbes*. (Available online at http://www.forbes.com/sites/tomgroenfeldt/2015/01/29/lenddo-creates-credit-scores-using-social-media/)

Hand, D. J., & Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*. Series A (Statistics in Society), pp.523-541.

Hosmer Jr, D. W., & Lemeshow, S. 2004. *Applied logistic regression*. John Wiley & Sons.

Hsieh, N. C. 2005. Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications* (28:4), pp. 655-665.

Huang, C. L., Chen, M. C., & Wang, C. J. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications* (33:4), pp.847-856.

Huang, J., & Ling, C. X. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* (17:3), pp.299-310.

Jensen, H. L. 1992. Using neural networks for credit scoring. *Managerial finance* (18:6), pp.15-26.

Ong, C. S., Huang, J. J., & Tzeng, G. H. 2005. Building credit scoring models using genetic programming. *Expert Systems with Applications* (29:1), pp.41-47.

Orgler, Y. E. 1970. A credit scoring model for commercial loans. *Journal of Money, Credit and Banking*, (2:4), pp. 435-445.

Karlan, D. S. 2005. Using experimental economics to measure social capital and predict financial decisions. *American Economic Review* (95:5), pp. 1688-1699.

Karlan, D. S. 2007. Social connections and group banking. *The Economic Journal* (117:517), F52-F84.

Lazarsfeld, P. F. & Merton, R. K. 1954.*Friendship as social process*. In M. Berger, T. Abel, and C. Page (Eds.), Freedom and Control in Modern Society. New York: Octagon.

Lee, L. F. 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics* (140:2), pp. 333-374.

Lee, T. S., & Chen, I. F. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* (28:4), pp.743-752.

Lewis, E. M. 1992. *An introduction to credit scoring*. Fair, Isaac and Company.

Mays, E. 2001. *Handbook of credit scoring*. Global Professional Publishi.

McPherson, M., Smith-Lovin, L. & Cook, J.M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (27), pp. 415–444.

Pennacchiotti, M. and Popescu, A.M., 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 430-438.

Poon, M. 2007. Scorecards as devices for consumer credit: the case of Fair, Isaac & Company Incorporated. *The sociological review* (55:s2), pp. 284-306.

Rosenberg, E., & Gleit, A. 1994. Quantitative methods in credit management: a survey. *Operations research* (42:4), pp. 589-613.

Shalizi, C. R., & Thomas, A. C. 2011.Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research* (40:2), pp.211-239.

Simon, R. December 16, 2015. Kiva Sets New Rules for U.S. Borrowers to Get Crowdfunded Loans. *The Wall Street Journal*. (Available at http://www.wsj.com/articles/kiva-sets-new-rules-for-u-s-borrowers-to-get-crowd-funded-loans-1450305877)

Srinivasan, V., & Kim, Y. H. 1987. Credit granting: A comparative analysis of classification procedures. *Journal of Finance* (42:3), pp. 665-681.

Strobl, C., Malley, J., & Tutz, G. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* (14:4), pp. 323.

Thomas, L. C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting* (16:2), pp. 149-172.

Thomas, L. C., Edelman, D. B., & Crook, J. N. 2002. *Credit scoring and its applications*. Siam.

Thomas, L. C., Oliver, R. W., & Hand, D. J. 2005. A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society* (56:9), pp. 1006-1015.

Van den Bulte, C. and Stefan W. 2007, *Social Networks and Marketing*. Cambridge, MA: Marketing Science Institute.

Wasserman, S., & Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge university press.

Wei, Y., Yildirim, P., Van den Bulte, C., & Dellarocas, C. 2016.Credit Scoring with Social Network Data. *Marketing Science* (35:2), pp. 234-258.

Weinland, D. & Robertson, B. November 17, 2014. Firms ride boom in credit profile services. *South China Morning Post*. (Available online at http://www.scmp.com/business/banking-finance/article/1641591/firms-ride-boom-credit-profile-services)

Wiginton, J. C. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis* (15:3), pp.757-770.

World Bank. 2015. Getting Credit. *The World Bank*. (Available online at http://www.doingbusiness.org/data/exploretopics/getting-credit)

Wu, X., Feng, Z., Fan, W., Gao, J. and Yu, Y., 2013. Detecting marionette microblog users for improved information credibility. In *Machine Learning and Knowledge Discovery in Databases,* pp. 483-498.

Yobas, M. B., Crook, J. N., & Ross, P. 2000. Credit scoring using neural and evolutionary techniques. IMA *Journal of Management Mathematics* (11:2), pp. 111-125.

Yoganarasimhan, H. 2012. Impact of social network structure on content propagation: A study using YouTube data. *Quantitative Marketing and Economics* (10:1), pp.111-150.

Yunus, M. 2007. *Grameen Bank at a glance*. Grameen Bank.

Zhu, X., Ghahramani, Z., & Lafferty, J. 2003. Semi-supervised learning using gaussian fields and harmonic functions. *In ICML* (3), pp. 912-919.

Zhu, X., & Goldberg, A. 2009. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.